Automated Information Extraction and Analysis for Information Synthesis

Catherine Blake, MComp, MS
Information and Computer Science
University of California,
Irvine
cblake@ics.uci.edu

Wanda Pratt, Ph.D.
Information School and
Division of Biomedical &
Health Informatics
University of Washington
wpratt@u.washington.edu

Tammy Tengs, Sc.D
Health Priorities Research Group
School of Social Ecology
University of California,
Irvine
tengs@uci.edu

Introduction

As the quantity of biomedical studies continues to increase, scientists in public health and biomedicine struggle to keep up with new findings, even in narrow areas of expertise. Our goal is to understand how scientists currently use the biomedical literature to answer research questions, and to develop tools to assist with that process. We observed that although advances in information retrieval have eased the task of identifying relevant articles, the challenge of aggregating information from within the retrieved set of articles remains. We call the iterative, collaborative process used by scientists to retrieve articles, extract information and then analyze the extracted facts, **information synthesis**[1].

Scientific articles in medicine and public health are comprised of two types of information. Primary information is the specific research question explored in an article, usually indicated by the title. Secondary information provides the context of a study and includes the age of subjects, demographics and behavioral habits. Scientists and clinicians struggle to integrate the primary results of a study into their work and have little time to utilize secondary information. For example, a scientist who studies factors that increase breast cancer risk has access to 3,900 articles each year, in addition to the more than 74,000 studies already in MEDLINE. Although factors reported in each of these articles could be associated with increased breast cancer risk, it is challenging to identify the studies that report a candidate factor as secondary information because indices are based on the primary purpose of a study. It is challenging to quantify the increased risk because the manual techniques employed to extract secondary information are time consuming and tedious, thus secondary information is a vastly under-utilized resource.

Methods

We are implementing a computer system to automatically extract secondary information from biomedical literature. Specifically, the system must extract (1) study information, such as the number of subjects with breast cancer and the geographic location of the study; (2) population information, such as the subjects' age, gender and ethnicity; (3) risk factor information, such the quantity and length of exposure, and (4) medical condition information, such as the location and severity of the disease. The study and population information are used to identify a similar population from the Behavioral Risk Factors Surveillance System (BRFSS) (www.cdc.gov/ccdphp/brfss), the world's largest telephone survey. The system will then use meta-analytic techniques to compare the risk-factor exposure rates reported in the literature with exposure rates reported in the BRFSS (see figure 1). We have implemented algorithms to extract the age and number of subjects with breast cancer, their tobacco usage, and the geographical location and time-frame of each article and the random-effects meta-analysis.

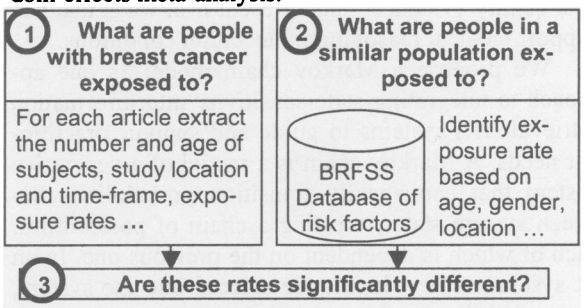

Figure 1 – Using biomedical literature to quantify the risk of breast cancer associated with tobacco usage.

Future Work

This work is the first step towards developing techniques to synthesize information from multiple articles. Our goal is to develop technology that supports the extraction and analysis phases of the information synthesis process that we observed, as scientists used biomedical literature to answer research questions. We are in the process of quantifying our existing extraction algorithms and developing new techniques to extract medical condition information.

Acknowledgements

A Multiple Investigator grant from the University of California, Irvine, funded portions of this work.

References

1. Blake, C. and W. Pratt. *Collaborative Information Synthesis*, *ASIST 2002* (In Press) Philadelphia.